DOCUMENT RESUME

ED 335 002 IR 015 169

AUTHOR Orey, Michael A.; And Others
TITLE Evaluation of Intelligent Tutoring Systems: What's Been Done and What Can Be Done.
PUB DATE 91
NOTE 18p.; In: Proceedings of Selected Research Presentations at the Annual Convention of the Association for Educational Communications and Technology; see IR 015 132.
PUB TYPE Information Analyses (070) -- Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Computer Assisted Instruction; Elementary Secondary Education; *Evaluation Methods; Higher Education; *Instructional Effectiveness; Instructional Systems; Intermode Differences; Mathematics Instruction; Microcomputers; Programing
IDENTIFIERS *Intelligent Tutoring Systems

ABSTRACT

This paper explores the realm of evaluation within the domain of Intelligent Tutoring Systems (ITS). Toward that end, Littman and Soloway (1988) described two facets of evaluation: external and internal. Described here are a variety of procedures that they did not address and some that they discounted. One of the procedures not explored was the use of Bloom's (1984) "2-Sigma Problem." Anderson, Boyle, and Reiser (1985) used this procedure to externally evaluate the Lisp Tutor. Priest and Young (1988) offered a variety of procedures that can be used for internal evaluation. These procedures are presented and POSIT (an ITS for subtraction of whole numbers) is used as an example case to illustrate the evaluation procedures. (26 references) (Author/BBM)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED335002

Title:

# Evaluation of Intelligent Tutoring Systems: What's been Done and What can be Done

Authors:

Michael A. Orey
Peter J. Legree
Philip D. Gillis
Estelle Bloom

IR015169

BEST COPY AVAILABLE

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Michael R. Simonson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

**Abstract**

This paper explores the realm of evaluation within the domain of Intelligent Tutoring Systems (ITS). Toward that end, Littman and Soloway (1988) describe two facets of evaluation that are examined: external and internal evaluation. Described here are a variety of procedures that they did not address and some that they discounted. One of the procedures not explored is the use of Bloom's (1984) "2-Sigma Problem". Anderson, Boyle and Reiser (1985) used this procedure to externally evaluate the Lisp Tutor. Priest and Young (1988) offer a variety of procedures that can be used for internal evaluation. These procedures are presented and POSIT (an ITS for subtraction of whole numbers) is used as an example case to illustrate the evaluation procedures.

594 3

# EVALUATION OF INTELLIGENT TUTORING SYSTEMS: WHAT'S BEEN DONE AND WHAT CAN BE DONE

## Introduction

While there has been a great deal of expectation and a small amount of evidence for the benefits (Anderson, Boyle, & Reiser, 1985; Raghavan & Katz, 1989; Kurland, Granville, & MacLaughlin, 1990) of Intelligent Tutoring Systems (ITS), there is at least as much equivocating evidence (Center for the Study of Evaluation, 1986; Sleeman, Kelly, Martinak, Ward, and Moore, 1988; Sleeman, et.al., 1989; Stasz, 1988). The available data on these evaluations do not yet seem to support large scale expenditures for ITS implementation, but weak research methodology may be at the root of the problem. However, researchers are beginning to address the issue of how to evaluate ITS (Littman & Soloway, 1988; Priest & Young 1988). Littman and Soloway (1988) suggest that evaluation of ITS has two facets: external and internal evaluation. External evaluation addresses the question, "What is the educational impact of an ITS on students?" (p. 209). We suggest that external evaluation may be most appropriately conducted for ITS by comparing the effectiveness of the ITS to that of human tutors and regular classroom instruction, although there are some dangers in using these groups for comparisons (Clark, 1983; Hogan, 1958; Reeves, 1986).

Internal evaluation, on the other hand, addresses the question, "What is the relationship between the architecture of an ITS and its behavior?" (Littman & Soloway, 1988, p. 209). Besides architectural concerns that are common to more traditional forms of computer-based instruction (CBI) such as interface usability, components specific to an ITS need examination (e.g., student models, expert models, pedagogical models and knowledge representation). Priest and Young (1988) present four different ways of evaluating the student model within an ITS. The student model may be the most widely evaluated aspect of an ITS. Besides Priest and Young's suggestions for evaluating the student model, we suggest some ideas for evaluating the other components. In the end, a combined evaluation that rigorously tests the ITS internally and externally may begin to reveal or put an end to the great expectation for ITS.

In the following pages we describe each of these evaluation approaches and use POSIT (Orey & Burton, 1990) as a test ITS to exemplify these evaluation processes. Before we begin the description of evaluation procedures, it will help to describe the ITS used to demonstrate the procedure. In addition, a short description of the data collection is provided.

## Description of POSIT

POSIT is an ITS for the tutoring of whole-number subtraction. The design of this system (see, Orey & Burton, 1990, for a detailed description of the design) is based somewhat on Anderson's (1987) ACT* theory of learning. Within this theoretical orientation, it is assumed that the learning of a cognitive skill builds from declarative knowledge. In terms of subtraction, the algorithm to be learned is made of a set of declarative information. During the development of this skill, this knowledge becomes consolidated into the specific skill (in this case, proficiency with a subtraction algorithm develops). Declarative knowledge is provided to the learner in a text form when errors occur or when the learner asks for help. For example, this is a typical message after making the Error of Omission (see Table 1 for error types) - Decrement (#1 and #2 are variables that are bound during the execution of the program) "*You have to complete the borrow into the ones place by*

*taking away one of the tens. So, the correct value for this area is #1 You typed #2 Please enter the correct value.*" Although this is a very brief description of POSIT, it should suffice for the purposes of this paper. It is also necessary to provide a short description of the participants and the method used for the collection of data for this paper.

Table 1. *The four error categories and instances of errors in each category (numbers are ranked according to number of occurrences).*

| Errors of Omission | Errors of Commission |
|---|---|
| 1. Increment (add 10) | 5. Increment |
| 12. Decrement (subtract 1) | 4. Decrement |
| 13. Both | 11. Both |
| 7. Neither | |
| **Errors of Creation** | **Other Errors** |
| 8. smaller-from-larger | 2. Fact error |
| 10. 0-N=N | 6. Typing error |
| 9. 0-N=0 | 3. Leading 0 |
| 14. N-N=N | |
| 15. N-N=0 with regrouping from | |
| 16. M-S=0, with S>M | |

## Data Collection

Two separate studies have been conducted on POSIT. The first was carried out in the spring of 1989 in a small town in southwest Virginia. The second was conducted in the fall of 1990 in a lower middle class suburb of Michigan. The Virginia study was conducted for the purpose of internal evaluation and the Michigan study was conducted for the purpose of external evaluation. A description of each group of participants follows.

Twelve participants were used in the Virginia evaluation. These participants were close to being 9 years old. Only those students who were having difficulties in subtraction were used. The selection of these participants was based on their teacher's recommendation and their performance on a subtraction test developed by the DEBUGGY system and reported in Van Lehn (1982). Out of a pool of 28 students who were having difficulties in subtraction, twelve participants were selected, but only ten of these completed the study. The data was collected over a five week period of time where the participants used the system about 30 minutes a day (averaging about 6 hours of total instruction). POSIT saves a great deal of information about each interaction. These data were used to demonstrate the internal evaluation procedure.

In the Michigan evaluation the entire third grade was used. This constituted 38 possible participants. One arrived late for the study and was eliminated and a second demonstrated mastery for subtraction prior to the beginning of the study which totaled 36 participants. The average age was just over 8 years. These participants were split into three groups; human tutor, computer tutor, and regular class. This yielded 12 participants for each group. The study was conducted over a five week period and the participants spent approximately 30 minutes a day working on subtraction. These data were used to demonstrate the external evaluation procedure.

The major benefit of the Michigan study was to provide appropriate feedback to the learner. The University of Georgia, Wayne State University, Apple Computer, Inc. and Oakland Elementary School, Royal Oak, Michigan, collaborated to implement this model. Both universities agreed to co-direct this project, provide technical expertise, and monitor and evaluate the program. Apple provided ten Macintosh SE's for a lab environment as well as the necessary system and utilities software.

The principal at Oakland Elementary agreed to obtain parental permission for students to participate in the POSIT program, meet with participating teachers to facilitate scheduling of students in the study, collect and record research data on a daily basis, and facilitate that data between the various sites and the University of Georgia. An integral factor was the utilization of parent volunteers to assist in implementation of this exploratory study by monitoring the various student groups involved. Most importantly, provisions were made to assure that all students who were part of the POSIT program in any of the three groups (parent volunteer, computer group, and non-computer control group) were able to receive computer instruction in a variety of curricular areas.

## External Evaluations: What's been done

Although over twenty ITSs have been developed, very few of these systems have been externally evaluated. Many of these systems were developed on AI workstations that cannot be used to economically distribute training. These systems were primarily designed to demonstrate the types of interactions and approaches that could be used to deliver training. The fact that the these ITSs were not intended for widespread use may account for the paucity of formal evaluations.

Much ITS that is now being developed will run on microcomputers because of improvements in PC based software. These systems, although expensive to develop, have the potential to become cost-effective training systems because the computer programs will run on readily available microcomputers. However, because microcomputer based ITS remains expensive to develop, it is critical that performance data be collected to determine its effectiveness.

In the past, the developmental cost of ITS technology has been justified by the claim that this technology may approach the effectiveness of one-on-one human tutorial instruction (Anderson & Reiser, 1985). In fact, at least one source (Littman & Soloway, 1988) has suggested that ITS effectiveness will eventually surpass the effectiveness of the human tutor. This should have major implications for training and education because students who are taught by human tutors perform approximately two standard deviations better on instructional exams than students taught by traditional methods (Bloom, 1984). The demonstration that ITSs lead to similar gains in training effectiveness should also justify continued support for developing practical microcomputer based ITS.

Evaluating ITS technology with well designed research studies is critical to avoiding a repetition of the history of the evaluation of conventional computer aided instruction (CAI). Initial metareviews of CAI studies (Kulik, Bangert & Williams, 1983; Bangert-Downs, Kulik & Kulik, 1985; Kulik, Kulik & Shwalb, 1986) supported optimistic estimates of the effectiveness of CAI and reported effect size estimates ranging from 0.25 to 0.42. However, a metaanalysis (Clark, 1985), which eliminated poorly designed studies from Kulik and Kulik's reviews, estimated a much smaller effect size for CAI effectiveness, 0.09. This later estimate underscores the importance of using sound evaluation procedures to estimate the effectiveness of microcomputer based ITS.

*Evaluations of Limited ITSs*

Evaluations that have been performed on ITSs with limited scope have not been supportive of this technology. This problem is most apparent in the evaluations of older ITSs and probably reflects the fact that the first ITS applications were designed to demonstrate the instructional capabilities of ITSs. Less attention was paid to either the scope of the system, or the practical impact of the system. Performance evaluations of these systems have generally resulted in poor evaluation results. The evaluations of PROUST, PIXIE, and WEST exemplify these problems.

*PROUST*

PROUST was developed at Yale University to tutor students on two optional Pascal programming language problems over a semester course. The ITS was evaluated by comparing the midterm and final grades of students who opted to use PROUST for the optional homework assignments, with students who completed the homework assignments on their own. No differences were demonstrated between the two groups, although differences were demonstrated between students who completed the optional homework assignments versus those who did not (Center for the Study of Evaluation, 1986).

The major problem with the PROUST evaluation is that the ITS only supported two problems. It is unlikely that group differences would have been demonstrated had human tutors been used for the two optional assignments instead of the ITS. It is not reasonable to expect group differences supporting ITS technology given the scope of PROUST. The lack of empirical differences should not influence expectations concerning the effectiveness of ITS technology.

*Pixie*

Pixie, unlike WEST and PROUST, is an ITS that is still under development and will support a broad topic area, algebra instruction. Sleeman, Kelly, Martinak, Ward, and Moore (1988, 1989) describe a series of four experiments relating to the Pixie project. The Pixie research is relevant to this paper because it demonstrates that the effectiveness of a limited tutoring system is difficult to demonstrate with either computers or with human tutors.

The Pixie experiments compared the effectiveness of three tutoring strategies: Model Based Remediation (MBR), Reteaching, and a control approach. The same basic paradigm was used across the four studies, students were randomly assigned to groups of eight and were tutored according to one of the three methods for one week during their Algebra class. Thus the groups were tutored for less than five hours. The impact of the three tutoring strategies was assessed with a knowledge test following the last session.

The two experimental conditions differed in that MBR tailors instruction to the needs of the individual student while Reteaching simply reiterates the information to the student following an error. Thus MBR is ideally suited as an ITS strategy while Reteaching corresponds to a CAI approach. Students in the control condition were informed when errors were made, but no remediation was attempted.

The first study compared the effectiveness of the three approaches when presented via computers. The remaining three studies estimated the effectiveness of the three approaches when human tutors were used. The results for the human tutors and the computers were consistent in that both MBR and reteaching led to significant improvements in performance relative to the control condition, however significant differences were never demonstrated between the MBR and Reteaching conditions.

The PIXIE research indicates that five high school class periods of exposure to two tutoring approaches cannot be used to compare the effectiveness of the two approaches. This was true for instruction delivered via human tutors as well as the ITS. If the Pixie evaluation studies had used larger sample sizes, then group differences might have been demonstrated. However, effect size estimates would probably still be small.

*WEST*

The WEST ITS helps students play the game, How the West Was Won, which was designed to teach operator precedence. WEST's impact on learning is limited by the effectiveness of the game for teaching operator precedence and by the exposure of students to the tutor, which is less than one hour. WEST was evaluated by comparing the performance of three small groups of subjects (5 to 7 subjects per group) on mathematics tests one day after using the program for one session (Center for the Study of Evaluation, 1986).

The evaluation did not support the ITS approach in that group differences were not demonstrated. However, it is not surprising that the system did not have a measurable effect on performance because one hour of tutoring cannot be expected to have a large impact on performance. We have not found any data suggesting that one hour of tutoring, in comparison to one hour of traditional instruction, can have a large impact on performance. Thus the WEST evaluation findings are not useful in estimating the effectiveness of ITS technology and are damaging to the ITS field because they suggest that ITSs are not effective.

*Evaluations of Well Developed ITSs*

In contrast to ITSs designed with limited scope, ITSs developed for practical use tend to be positively evaluated. The MACH III, Lisp ITS, and Smithtown exemplify well developed systems. All three tutors were designed for an actual training requirement and cover a large amount of course material. The MACH III supports 32 hours of Army training, the Lisp ITS covers a one semester course at Carnegie Mellon University, and Smithtown corresponds to approximately one-third of an economics course.

All three of these ITSs have been favorably evaluated by the demonstration of statistically significant group differences. Furthermore, the group differences were meaningful in that each of the tutors had a substantial effect on course performance. The major problem with these studies was that the authors were not consistent in their reporting of effect size estimates or other statistics.

*MACH III*

The MACH III ITS evaluation report includes mean and variance estimates for students taught with the ITS, versus those taught with traditional classroom instruction. The ITS was used for approximately 32 hours of instruction and the evaluation data indicate an effect size of approximately one standard deviation (Kurland, Granville, & MacLaughlin, 1990).

*Lisp ITS*

The Lisp ITS was evaluated twice (Anderson & Reiser, 1985; Anderson, Boyle & Reiser, 1985). In the first evaluation, groups of 10 students learned Lisp with either a human tutor, the Lisp ITS, or through self instruction. During the first evaluation, the Lisp ITS covered 6 of the 16 chapters covered in the course, or 38 percent of the course. The first evaluation did not demonstrate group differences on course exams, however, time data indicate that the human tutor, ITS, and traditional instruction groups required different amounts of time, 11.4, 15, and 26.5 hours, to cover the same material.

In the second Lisp ITS evaluation, two groups of ten students learned Lisp with either the Lisp ITS or traditional classroom instruction. During this evaluation, the ITS covered 9 of the 16 chapters, or 56 percent of the course. A class exam indicated that the Lisp ITS students "performed 43 percent better than" the traditional classroom instruction group.

Unfortunately variance estimates were not reported for either Lisp ITS evaluation, thus it is not possible to estimate effect size. However, the time estimates and the classroom test scores are consistent with the 1.0 effect size estimated for the MACH III. The fact that differences were only demonstrated on the knowledge test for the second Lisp evaluation is most consistent with the assertion that ITS evaluations are more likely to be successful for better developed ITSs.

*Smithtown*

Two Smithtown evaluations are described by Raghavan and Katz (1989). The first Smithtown evaluation demonstrated that students who worked with the tutor spent much less time, 5 hours, to learn material that required 12 hours of time for students who did not have access to the tutor. This time savings estimate is consistent with time data from the Lisp ITS evaluations.

A second evaluation study demonstrated a 15 point advantage on a course test for students who had access to Smithtown during instruction as compared to students who did not. These results are consistent with the Lisp ITS data, as well as with the 1.0 standard deviation effect size estimated for the MACH III

## External Evaluation: What should be done

These studies were analyzed to identify promising external evaluation procedures and criteria. The review indicated the need for three criteria that are particularly relevant to ITS evaluation studies. First, ITS evaluations should utilize three groups of students: the ITS group, a traditional instruction group, and a human tutor group. Second, only well developed ITSs should be evaluated to estimate the impact of this technology on learning. Third, ITS evaluations should utilize larger group sizes than have been used in the past. We suggest that ITS researchers ought to adopt these criteria for future ITS evaluations.

### *Required Groups : ITS, Human Tutor, & Traditional Instruction*

Logic dictates that a minimum of three groups of subjects should be included in an ITS evaluation: a traditional classroom control, a human tutorial, and an ITS group. Although most ITS evaluations have not included these three groups, all three are critical in order to insure meaningful results. Including all three of these groups is important for a number of reasons.

ITS proponents have consistently justified ITS development on the claim that human tutoring is a highly effective means of instruction (Anderson & Reiser, 1985; Littman & Soloway, 1988). However, not all learning objectives can be assumed to be most effectively taught in a one-on-one environment. Comparing the effectiveness of a human tutor and a traditional instruction group would assess this assumption for the ITS subject matter and could justify basing ITS instruction on models of human tutoring. If human tutors cannot be demonstrated to be more effective than classroom instruction for a specific learning objective, then it is not logical to construct a system that emulates human tutors. (Obviously, it is preferable for this comparison to be made prior to developing the ITS.)

If the assumption can be made that human tutoring is more effective than traditional instruction, then including both controls in the evaluation allows the power of the research design to be assessed. Under this assumption, the experimental finding that these two groups do not differ is a Type I error and

indicates problems with the experimental design. This finding could occur if the evaluation design used small sample sizes, or inaccurate performance measures.

If the human tutor and the traditional instruction groups differ, then the magnitude of the difference can be used to estimate the magnitude of the differences that can be expected between the ITS and traditional instruction groups. This information may be used to assess the adequacy of the sample sizes and performance measures used in the evaluation. This is particularly important if the evaluation does not demonstrate differences favoring the ITS group.

It is also important to compare the effectiveness of the ITS approach to the effectiveness of human tutors to allow an estimation of the extent to which the salient aspects of human tutor interactions have been encoded into the ITS. It can be expected that as ITS technologies improve, computer based tutors will approximate the effectiveness of human tutors. This comparison allows ITS developers to estimate the quality of their programs in relation to human tutor effectiveness and determine when other models are needed to further improve the effectiveness of the ITS.

Evaluating the cost-effectiveness of an ITS requires the ITS group to be compared to a traditional classroom control because in most cases, this is the least expensive instructional alternative. If the ITS cannot be shown to be either more effective than traditional instruction, or equally effective and less costly than traditional instruction, then it is not logical to develop these systems for practical applications.

### *Impact of the System*

ITS evaluations have assessed the impact of both well developed and limited ITSs on learning. Within these evaluations, most of the well developed ITSs have been favorably evaluated, while the evaluations of limited systems have not supported ITS technology. Well developed systems are defined as systems that can be expected to have a large impact on performance, while a small impact on performance can be anticipated for a limited systems.

The magnitude of the impact of a system on performance is important for two reasons. First, demonstrating a small effect requires much greater evaluation resources than are needed to demonstrate a large effect. These resources may not always be available. Second, whether or not adequate evaluation resources are available, ITSs must be well developed in order for an evaluation to indicate a large impact on performance. This is because effect size is independent of sample size and is calculated as the ratio between the mean difference between treatment groups to the standard deviation of the observations: (Mean Difference)/(sd). This is important because ITSs that have only small effects on performance cannot be used to justify continued research and development in ITS technology. It follows that ITS evaluations should focus on well developed tutors.

POSIT was evaluated in this way in the Michigan study. The regular class (mean=358.714) was compare to the ITS (mean=296.286) and the combined standard deviation for these two groups was 79.618 with the dependent measure being the time until mastery in minutes. The resulting effect size is 0.784 (358.714-296.286/79.618). This figure is close that indicated in the MACH III evaluation. The reason for it being less than one may be due to the fact that the experiment was prematurely terminated. Because only seven of the twelve participants in the regular class instruction achieved mastery within the time scheduled for the experiment, the first seven to achieve mastery in the ITS group were used for this analysis (all twelve had achieved mastery within the scheduled time).

### *Group Size Requirement : Power Analyses*



BEST COPY AVAILABLE

One problem with many evaluations is the use of small groups of subjects. Table 2 contains a listing of sample sizes for evaluations of nine ITSs. The outcome of the evaluation is also summarized in Table 2. Table 2 indicates that all of the studies used sample sizes that were less than 20. This is problematic from the standpoint of power analyses because small group sizes result in very weak ITS evaluation designs. Much larger numbers of subjects need to be included in ITS evaluation designs in order to accurately assess the impact of ITS technology.

The data summarized in Table 2 demonstrate that the issues of group size and effect size are closely linked to the probability that the system is likely to have a demonstrable effect on performance. The only evaluation studies that were favorable were those that utilized larger samples and those that have already been identified as having a substantial effect on performance (Refer to Table 2).

Table 2. *Sample sizes used in ITS evaluations.*

| ITS | Sample Size Per Group | Number of Conditions | Referencea | Outcomeb |
|---|---|---|---|---|
| Lisp ITS | 10 | 3 | 1,2 | 1 |
| Lisp ITS | 10 | 2 | 1,2 | 1 |
| PROUST | 2-19 | 2c | 3 | 0 |
| WEST | 5-7 | 3 | 3 | 0 |
| MACH III | 11 | 2 | 4 | 1 |
| Smithtown | 15d | 2 | 5 | 1 |
| Smithtown | 15 | 3 | 5 | 1 |
| Geometry Proofs | 3e | 1 | 6 | 0 |
| Pixie | 8d | 3 | 7,8 | 0 |
| Pixie | 8d | 3 | 7,8 | 0 |
| Pixie | 6 | 4 | 7,8 | 0 |
| Pixie | 8-9 | 3 | 7,8 | 0 |
| ADCS | 2 | 3 | 9 | 0 |
| POSIT | 7f | 3 | 10 | 1 |
| Mean | 8.1 | 2.6 | | 0.43 |

a: Refer to reference list for Table 2.
b: "1" indicates a favorable evaluation, "0" indicates no demonstrated group differences.
c: Four levels within each condition.
d: Only the total number of subjects was reported. Table values are estimates
e: Three students were pilot tested on the ITS.
f: Although the experiment began with 12 per group, it ended with only 7

References

1. Anderson & Reiser, 1985;
2. Anderson, Boyle, & Reiser, 1985
3. Center for the Study of Evaluation, 1986
4. Kurland, Granville, & MacLaughlin, 1990
5. Raghavan & Katz, 1989
6. Schofield & Evans-Rhodes, 1989
7. Sleeman, Kelly, Martinak, Ward & Moore, 1988
8. Sleeman, Kelly, Martinak, Ward & Moore, 1989
9. Crooks, Kuppin & Freedy, 1978
10. Orey, 1990

Another way to view this relationship is to consider the relationship between sample size, power, alpha probability level, and effect size. In general, as either the effect size of a treatment or the sample sizes used in a study increases, then the power of the study will also increase. Cohen (1977) has calculated power estimates for various effect sizes and sample sizes. A portion of the values reported by Cohen are listed in Table 3.

Assuming an effect size of 1.0 standard deviation unit, which is the value estimated for the MACH III effect size and is consistent with the Lisp ITS and Smithtown Effect Size, Table 3 indicates that the sample sizes used in these three successful evaluations were marginally adequate. The power associated with the studies is approximately equal to .55, .60, and .83 for the sample sizes used in the Lisp ITS, MACH III, and Smithtown evaluations. In fact, the designers of these evaluations were fortunate to have demonstrated group differences in these studies.

Table 3. *Power analysis estimates: Required sample size per group at the alpha=.05 (2-tail) level by effect size and power.*

| Power | Effect Size=\|MeanA-MeanB\|/sigma | | | | | | |
|---|---|---|---|---|---|---|---|
| | .20 | .40 | .60 | .80 | 1.00 | 1.20 | 1.40 |
| .25 | 84 | 22 | 10 | 6 | 5 | 4 | 3 |
| .50 | 193 | 49 | 22 | 13 | 9 | 7 | 5 |
| .60 | 246 | 62 | 28 | 16 | 11 | 8 | 6 |
| .80 | 393 | 99 | 45 | 26 | 17 | 12 | 9 |
| .95 | 651 | 163 | 73 | 42 | 27 | 19 | 14 |
| .99 | 920 | 231 | 103 | 58 | 38 | 27 | 20 |

The impact of effect size on power is also demonstrated by comparisons within the MACH III and Lisp ITS evaluations. The MACH III evaluation compared experimental and control groups on four separate exams. Three of the exams were related to a small portion of the ITS or were psychometrically poor; group differences were not observed for these scales. On the fourth exam, the groups were significantly different at the p=.04. Thus the MACH III evaluation nearly led to a finding of no significant group differences.

The Lisp ITS was evaluated twice, the ITS covered 38 percent of the course for the first evaluation and 56 percent of the course during the second evaluation. Although the ITS groups in both evaluations spent less time solving problems, the ITS group performed significantly better than a control group only on the knowledge exam of the second evaluation. In the first evaluation, group differences were not demonstrated on a knowledge exam. It follows that given the small sample sizes used in the Lisp evaluations, the effect of the ITS on student performance could have been ea. ily missed.

These considerations underscore the importance of evaluating well designed ITSs with large sample sizes. A conservative expectation for the purposes of power analyses is to assume an effect size of .80. Then it can be estimated that sample sizes greater than 42 per group are needed to help ensure that the ITS evaluation will not result in a Type 2 error, i.e. missing a real effect, at the 95 percent probability level (Refer to Table 3).

## *Other Considerations*

*Outcome measures*

Ideally, performance data should measure both time savings, or power (percent correct). Unfortunately, these two categories conflict with each other because they are negatively correlated within individuals, i.e. the faster a person responds, the less likely that response is to be correct. Past ITS (and CAI) evaluations have used both types of measures, but the emphasis has usually been on power scales.

The strategy of attempting to demonstrate both time savings and test improvement can be problematic because the effect of the ITS is split across several dimensions. (Whereas by holding one dimension constant, e.g. time, the treatment effect would primarily affect the other dimension, i.e. percent correct.) Once the effect is split across several dimensions, demonstrating significant improvements due to the ITS on any one dimension may require a larger sample size. For this reason it is usually logical to limit variations in time on the ITS and concentrate on demonstrating differences in power scales.

Ideally, the tests should estimate the effectiveness of the ITS across several points of generalization. For example, the outcome measures should access performance on problems specifically related to the tutor, as well as on general class exams. This strategy helps to ensure that the evaluation will assess the overall effectiveness of the tutor and yield interpretable patterns of results.

*Reporting Statistics*

One shortcoming with many ITS evaluations is that the tests are not adequately described. Neither reliability, validity, effect size, nor variance estimates are reported. Covariates are often neither included nor discussed in the evaluation reports. Without these values it is difficult to assess the meaning of the evaluation. This is particularly true when the groups fail to be significantly different. In this case it is critical that the power of the research design be estimated.

*Cognitive Science Approach*

Littman and Soloway (1988) suggest that since ITS development is embedded in the field of cognitive science, a cognitive science perspective on evaluation should be taken. By a cognitive science perspective, it is implied that the system attempts to understand the cognitive processes and/or the knowledge structures (states) of the learner. Therefore, the evaluation should follow these lines. That is, the evaluation should focus on how the student is performing some process or how the student has organized some knowledge rather than on a number correct as measured on a typical achievement test. A general achievement test does not get at the level of specificity that an ITS can examine. Hence, learning is viewed as a continuous process and the system would be evaluated in terms of its ability to recognize a learner's current state and how well the learner is helped to progress from the current state to an advanced state.

However, gross analyses, as was described earlier, are also useful if they are well designed. Much of the content of the ITS can be measured on a test that was developed for the evaluation of the content area of the ITS. The processes that learners are learning from the ITS can manifest themselves both in terms of the actual knowledge and processes, but also in the product of these processes (i.e., answers on an achievement test). Therefore, there should be a distinction made between these two levels of external evaluation. At one level, the ITS is evaluated in terms of the knowledge states and cognitive processes which are a part of the design of the ITS. At the other level, the ITS is evaluated based upon performance

on an achievement test specific to the domain of the ITS (as was done with POSIT in the Detroit study to determine mastery).

Table 4. *The frequency of errors in each error category on the first and last days of the treatment in Virginia.*

| **Errors of Omission** | **First** | **Last** | **Errors of Commission** | **First** | **Last** |
|---|---|---|---|---|---|
| Increment (add 10) | | | Increment | | 1 |
| Decrement (subtract 1) | 11 | 1 | Decrement | 5 | 1 |
| Both | | | Both | | |
| Neither | | 1 | | | |
| **Errors of Creation** | | | **Other Errors** | | |
| smaller-from-larger | 1 | | Fact error | 1 | |
| 0-N=N | | | Typing error | 5 | 2 |
| 0-N=0 | 1 | | Leading 0 | 13 | 1 |
| N-N=N | | | | | |
| N-N=0 with regrouping from M-S=0, with S>M | | | Incorrect diag. | 6 | 6 |

In terms of POSIT, the external evaluation at the first level would be in terms of the specific declarative knowledge (productions) that make up POSIT. Operationally, this means that there should be a reduction in the number of errors that POSIT identifies. That is, POSIT should be helping students eliminate errors in their algorithms. Table 4 shows the number of errors on the first and last days of the field test made by the ten participants that used POSIT. The first result of importance is the fact that there was an overall reduction in the number of errors (from 30 to 13 errors). However, on a cognitive science level, the results indicate that the specific error productions that the participants were using at the beginning of the field test were greatly reduced by the end of the field test. The most pronounced is the least significant -- the leading zero error type. This result is not surprising since the regular classroom teacher taught the students to fill in all spaces, so leading zeros were correct in the classroom. On the other hand, Errors of Omission were nearly as frequent as the leading zero error. This decrease would indicate that the children's procedures were being modified by their having used POSIT. Another approach might be to examine the number of steps that were carried out mentally and the time it takes between steps (this data is available, but not analyzed).

## Internal Evaluation

Internal evaluation is more like research than it is like evaluation. Besides testing the functionality of the system, the theoretical assumptions about the manner in which learners learn is being tested. Internal evaluation, according to Littman and Soloway (1988), is the evaluation of the relationship between the architecture and the behavior of the ITS. The point of this evaluation is to maintain the goals of the tutor as it goes through its many revisions. A somewhat more powerful evaluation procedure was suggested by Priest and Young (1988) which might be used for internal evaluation. This procedure revolves around a cognitive science evaluation procedure in that it evaluates the internal theoretical power of the learner model.

Priest and Young (1988) suggest that there are four measures of the accuracy of the model of the learner. Those four measures are: 1) Percentage correct diagnosis; 2) Error fit measure; 3) Micro-theory evaluation quotient; and, 4) The cumulative hit curve. The first measure is the easiest, but there is no way to determine the true power of the learner model from this value. Afterall, the system may have an unique explanatory production for every error made (Young and Priest refer to this as a trivial case, which theoretically it is, but pragmatically it would be mind boggling to develop such a system). This does not contribute to a theoretical view of learning and therefore is not a powerful theory, yet the percentage correct would be 100 percent. Notice that in Figure 1 POSIT has a percentage correct of 76 percent.

The second measure, the error fit measure, is especially sensitive to false error analyses. Obviously, by subtracting wrong predictions, this value will decrease. The lower the number, the less viable the theory. Notice in Figure 1 that POSIT has a value of 0.665. The third measure specifically addresses the problem of a trivial theory. To ensure that the theory is robust enough, it is necessary to subtract the number of productions (or variant procedures) from the number of correct diagnoses. In this way, each production must be powerful enough to explain many errors or conditions. Perhaps the most powerful of the evaluation methods is the fourth: the cumulative hit curve. This curve plots the percentage of errors explained versus the number of productions used. The straight line represents the trivial case (1 to 1 correspondence between errors and productions). Figure 1 shows the cumulative hit curve for the first stage of evaluation of POSIT.

These evaluation measures help evaluate the theoretical power of a learner model within the ITS. Further, we can judge our progress by showing how the new system out performs the old system. We can use an excepted guideline. However, it is not readily apparent whether POSIT is good or bad based on these numbers. What is needed is for all systems to report these numbers so that an industry standard can be established.

## Conclusions

In conclusion, no evaluation procedure is complete without some if not all of the procedures described in this paper. It is quite important to perform evaluations on an ITS both internally and externally. The external evaluation seems to be more important with regards overall instructional effectiveness while the internal evaluation is concerned with theoretical power. ITS affords the opportunity to explore theoretical issues within cognitive science as well as instructional issues within the a cognitive instructional science.

*Figure 1.* Priest and Young's (1988) four measures of internal theoretical power and their values for POSIT.

1) Percentage correct diagnosis.

$$\frac{\text{Number of correct diagnoses X 100}}{\text{Number of data points}} = \frac{371 \text{ X } 100}{489} = 76\% \text{ for POSIT}$$

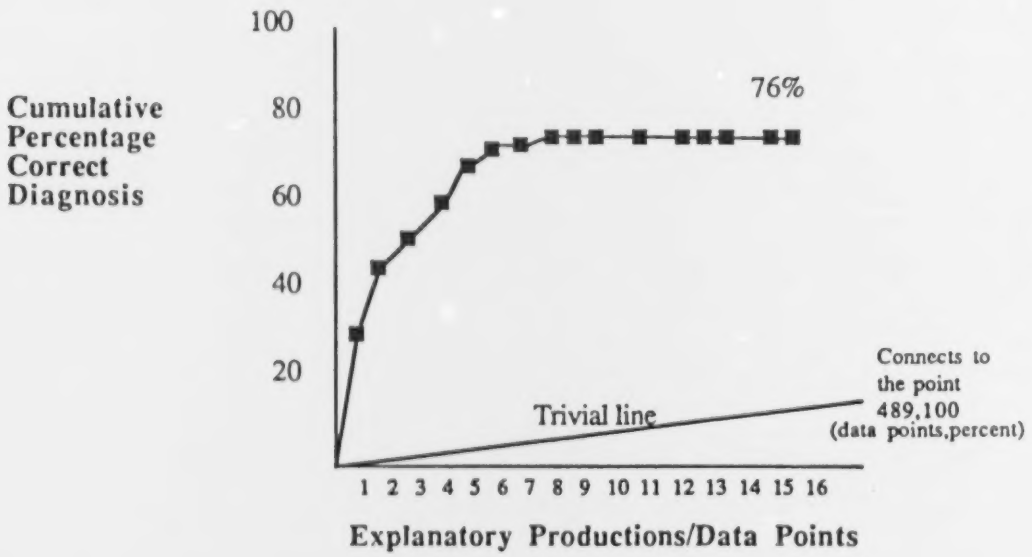
2) Error fit measure.

$$\frac{\text{Number correct diag. - Number wrong predictions}}{\text{Number of data points}} = \frac{371 - 46}{489} = 0.665 \text{ for POSIT}$$

3) Micro-theory evaluation quotient.

$$\mu = \frac{\text{Number correct diag. - Number of productions}}{\text{Number of data points}} = \frac{371 - 16}{489} = 0.726 \text{ for POSIT}$$

4) The cumulative hit curve.



## References

Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem solutions. *Psychological Review. 94*(2), 192-210.

Anderson, J. R., Boyle, C. F., & Reiser, B. J. (1985). Intelligent tutoring systems. *Science. 228*, 456-462.

Anderson, J. R. & Reiser, B. J. (1985). The Lisp tutor. *Byte,10* (4), 159-175.

Bangert-Downs, R. L., Kulik, J. A. & Kulik, C. C. (1985).Effectiveness of computer based education in secondary schools. *Journal of Computer Based Instruction, 12*, 59-68.

Bloom, B. S. (1984). The 2-sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4-16.

Center for the Study of Evaluation, University of California, Los Angelos. (1986). Intelligent computer aided instruction (ICAI): Formative evaluation of two systems (Army Research Institute Research Note 86-29). Alexandria, VA: U.S. Army Research Institute. (DTIC No. AD-A167 910).

Clark, R.E. (1983). Reconsidering research on learning from media. Review of *Educational Research, 53*, 445-459.

Clark, R. E. (1985). Evidence for confounding in computer-based instruction studies: Analyzing the meta-analyses. *Educational Communication Technology Journal, 33*, 249-262.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York, NY: Academic Press.

Crooks, W. H., Kuppin M. A. & Freedy, A. (1978). Application of adaptive aiding systems to computer-assisted instruction: Adaptive computer training system (ACTS) (Army Research Institute Technical Report 78-A6). Alexandria, VA: U.S. Army Research Institute. (DTIC No. AD-A056 900).

Hoban, C.F. (1958). Research on media. *AV Communications Newsletter, 6*(3), 169-178.

Kulik, J. A., Bangert, R. L. & Williams, G. W. (1983). Effects of computer-based teaching on secondary school students. *Journal of Educational Psychology, 75*, 18-26.

Kulik, C. C., Kulik, J. A. & Shwalb, B. J. (1986). The effectiveness of computer based adult education: A meta-analysis. *Journal of Educational Computing Research, 2*, 235-252.

Kurland, L. C., Granville, R. A., & MacLaughlin, D. M. (1990).*Design, development, and implementation of an intelligent tutoring system (ITS) for training radar mechanics to troubleshoot*. Unpublished manuscript, Bolt

Beranek, and Newman, Systems and Technologies Corporation, Cambridge, MA.

Littman, D., & Soloway, E. (1988). Evaluating ITSs: The cognitive science perspective. In M. C. Polson and J. J. Richardson (Eds.), *Foundations of Intelligent Tutoring Systems*, (pp.209 - 241) Hillsdale, NJ: Lawrence Erlbaum.

Orey, M. A, (1990). *Results of external evaluation.* Unpublished data.

Orey, M. A. (1989). *POSIT: Process Oriented Subtraction-Interface for Tutoring*. Unpublished dissertation, VPI & SU, Blacksburg, VA.

Orey, M. A., & Burton, J. K. (1990). POSIT: Process oriented-subtraction interface for tutoring. *Journal of Artificial Intelligence in Education, 1*(2), 77-104.

Priest, T., & Young, R. M. (1988). Methods of evaluating micro-theory systems. In John Self (Ed.), *Artificial Intelligence and Human Learning: Intelligent Computer-aided Instruction* (pp.124-137) London: Chapman and Hall Computing.

Raghavan, K. & Katz, A. (1989). Smithtown: An Intelligent Tutoring System. *Technological Horizons in Education Journal, 17* (1), 50-54.

Reeves, T.C. (1986). Research and evaluation models for the study of interactive video. *Journal of Computer-Based Instruction, 13*(4), 102-106.

Schofield J. W., Evans-Rhodes, D. & Huber, B. R. (1989).Artificial intelligence in the classroom: The impact of a computer-based tutor on teachers and students (O.N.R. Technical Report No. 3). Arlington, VA: Office of Naval Research.

Sleeman, D., Kelly, A. E., Martinak, R. Ward, R. D., & Moore, J.L. (1988). Diagnosis and remediation in the context of intelligent tutoring systems (ARI Research Note 88-66). Alexandria, VA: U.S. Army Research Institute. (DTIC No. AD-A199 024).

Sleeman, D., Kelly, A. E., Martinak, R. Ward, R. D., & Moore, J.L. (1989). Studies in the diagnosis and remediation of high school algebra students. *Cognitive Science, 13*, 551-568.

Stasz, C. (1988, April). *An intelligent tutor for basic algebra: Preliminary data on student outcomes*. Paper presented at the annual meeting of AERA in New Orleans.

VanLehn, K. (1982). Bugs are not enough: Empirical studies of bugs, impasses, and repairs in procedural skills. *Journal of Mathematical Behavior, 3*, 3-72.